Automated Detection of Premature Atrial Contractions in Noisy Exercise Electrocardiograms

Anna M Bator[1], Stefan van Duijvenboden[1], Rohan S Wijesurendra[1,2], Aiden Doherty[1]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[2]Department of Cardiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom

Abstract

Exercise electrocardiography (ECG) provides clinically important information on rhythm disturbances occurring under physiological stress, but is challenged by motion artefacts, muscle noise, and baseline drift. Premature atrial contractions (PACs) are commonly observed during exercise and may unmask arrhythmic burden, but due to their similarity to normal sinus beats, reliable automated detection remains challenging, particularly from single-lead recordings.

We developed a CNN–BiLSTM U-Net with self-attention to detect PACs in single-lead exercise ECG signals from a UK Biobank exercise cohort. Performance was assessed on a held-out test set using beat-level metrics, and agreement in participant-level PAC counts.

On the held-out test set, the model achieved a beat-level PAC F1 score of 0.74, with precision of 0.80 and recall of 0.70 within a multi-class classification framework (overall macro F1 0.84). Bland–Altman analysis demonstrated good agreement between predicted and ground truth PAC counts per participant, with a mean bias of −0.14 beats and limits of agreement (–3.33, 3.05).

This work demonstrates that automated PAC detection in noisy single-lead exercise ECGs is feasible, enabling scalable, population-level phenotyping of ectopy under cardiac stress.

1. Introduction

Premature atrial contractions (PACs) are early heartbeats originating in the atria, and are one of the most common cardiac arrhythmias. Whilst often asymptomatic, they are highly prevalent, with up to 99% of healthy adults over the age of 50 presenting with at least one PAC per 24 hours [1]. Although traditionally considered benign, higher PAC burden in resting and ambulatory ECGs has been associated with incident atrial fibrillation and other adverse outcomes [2,3].

Unlike current literature's focus on Holter monitoring and standard 10-second electrocardiograms (ECGs) [2,3], we examine exercise testing ECGs, which capture the heart's response to stress and can unmask arrhythmias not present at rest [4]. However, manual ECG annotation at scale is time-consuming and prone to human error [5]. Machine learning models have been used successfully for automated heartbeat detection and classification tasks, but their targets have primarily been ambulatory and resting recordings [6,7] rather than exercise ECGs.

PACs represent a particularly difficult target for automated detection. Unlike premature ventricular contractions (PVCs), which produce distinctively wide QRS morphologies and large T-waves, PACs closely resemble normal sinus beats, making them the consistently worst-performing class in ECG classification benchmarks [6,8,9]. This is amplified in single-lead exercise recordings which are inherently noisy due to electrode motion and muscle artefacts [10] and often contain poorly visible P-waves, making heartbeat detection and classification particularly challenging [11,12]. To our knowledge, no validated method exists for automated PAC detection in exercise ECGs, representing an important gap that this study aims to address.

2. Methods

2.1. Dataset and annotation

The dataset was derived from the UK Biobank exercise cohort (N = 97,435), a subset of the population-based prospective cohort study of approximately 500,000 UK adults aged 40 to 69 [13] who were cleared on cardiovascular risk assessment and completed a submaximal cycle ergometer test with a single-lead ECG recording (15 s rest, 6 min cycling, 1 min recovery, workload individualised by risk and estimated fitness). Recordings with insufficient quality, bundle branch block, atrial fibrillation, or extreme heart rate values had already been excluded [14].

The training set (n = 695) was created by selecting participants with high beat-to-beat variability but no other major arrhythmias, enriching for ectopic activity to populate the minority classes (PAC, PVC). The held-out test set (n = 1442) was then sampled at random from the remaining participants and so retains a lower ectopic prevalence representative of the full UK Biobank cohort. During annotation, recordings were further excluded where the underlying rhythm was irregular or non-sinus such that prematurity could not be reliably assessed. Reported sample sizes are post-exclusion.

Ground truth labels were generated by two trained ECG annotators following predefined guidelines developed in consultation with an independent cardiologist, who adjudicated ambiguous cases. Beats that could not be annotated with sufficient confidence were labelled Unknown and masked out during both training and testing. Annotators reserved the Unknown class for beats that were genuinely uninterpretable; such beats accounted for 0.3% of the training set and <0.1% of the test set (Table 1).

Table 1. Distribution of ground truth labels across the training and test cohorts. The training set was selected for high PAC and PVC counts whilst the test set is a random sample from the full UK Biobank cohort.

	Training	Test
Recordings	695	1442
Total Beats	493,892	982,424
Normal	439,968 (89.1%)	978,837 (99.6%)
PAC	25,494 (5.2%)	1,753 (0.2%)
PVC	27,166 (5.5%)	1,758 (0.2%)
Unknown	1,264 (0.3%)	76 (<0.1%)

2.2. Pre-processing

Raw ECG signals, sampled at 500 Hz, were pre-processed using a standardised pipeline. This included a 50 Hz notch filter to remove powerline interference, followed by a 4th order Butterworth bandpass filter with cut-off frequencies 2 and 45 Hz. The signal was then downsampled to 250 Hz to reduce computational demand and segmented into 10-second non-overlapping windows. Ground truth labels were generated by extending R-peak annotations to span the approximate QRS duration (160 ms for PVCs, 80 ms for Normal and PAC QRS complexes). Recordings were clipped at (-5mV, 5mV) to truncate high-voltage spikes and motion artefacts.

2.3. Model architecture and training

We trained a 1D U-Net with convolutional layers and a bidirectional long short-term memory layer in the bottleneck (CNN-Bi-LSTM) with self-attention using 5-fold cross-validation. The model accepted the 10-second ECG windows and performed sample-wise classification across the four classes: No-QRS, Normal, PAC, and PVC.

To address the severe class imbalance, focal loss with alpha weighting was used. The weights were estimated on a deterministic 10% subsample of the training set for memory efficiency. They were then power smoothed with an exponent of $\beta = 0.45$ to moderate the most extreme values, yielding weights of 0.57, 1.41, 4.94, and 3.63 for No-QRS, Normal, PAC, and PVC respectively, and ensuring compatibility with the focal loss focusing parameter $\gamma = 3.0$.

We applied class-specific confidence thresholds optimised using out-of-fold cross-validation F1 scores (PAC: 0.48, PVC: 0.56) to correct for distribution shifts between training and test sets. To preserve precision, PAC and PVC predictions falling below these thresholds were demoted to Normal or No-QRS, whichever was more probable.

2.4. Evaluation

Evaluation used two inference passes (original and a 5-second shift). Retaining the central 50% of the combined predictions ensured at least 2.5 seconds of preceding context for almost every sample. Each contiguous run of non-zero samples was collapsed to a single QRS complex prediction, positioned at the midpoint of the run and assigned the majority class within it to smooth predicted samples into beat predictions. Beats shorter than 40 ms were filtered out as spurious predictions.

Although the U-Net infers labels on a sample-basis, exact sample-level agreement with the ground truth is not clinically informative or relevant. We therefore report results on a beat basis, in which predicted and ground-truth beats were matched one-to-one, with the closest pair matched first, whenever their centroids fell within ±37 samples (148 ms), corresponding to the 150 ms threshold recommended by ANSI and widely used in the literature [15–17]. Any unmatched ground-truth beat was counted as a false negative and any unmatched predicted beat was counted as a false positive.

To filter out muscle and electrode motion artefacts, we flagged windows exceeding the training set's 99th percentiles for amplitude standard deviation (≈0.29 mV) or the first derivative standard deviation (≈0.098 mV per sample). Recordings with under 4 minutes of usable (unflagged) data were excluded, leaving 1430 test and 683 cross-validation recordings; no recording was excluded on raw duration alone. Within the retained test recordings, the remaining flagged noisy windows (178 total, 0.28%) were masked out before calculating final metrics.

The final performance was assessed using beat-wise, class-wise F1 score, precision, and recall, as well as a Bland–Altman analysis of participant-level agreement between predicted and ground truth class counts on the held-out test set.

3. Results

On the held-out test set, the model achieved a beat-level PAC F1 score of 0.74 (precision 0.80, recall 0.70). PVC and Normal beat-level F1 scores were 0.77 and 1.00 respectively, and the overall macro F1 score across Normal, PAC, and PVC (No-QRS class excluded) was 0.84 (Figure 1, Table 2). A drop in performance across all three classes was observed from cross-validation to the held-out test set (cross-validation PAC F1 0.90 (±0.02); PVC F1 0.92 (±0.01); macro F1 0.94 (±0.01); mean ± SD across k=5 folds).

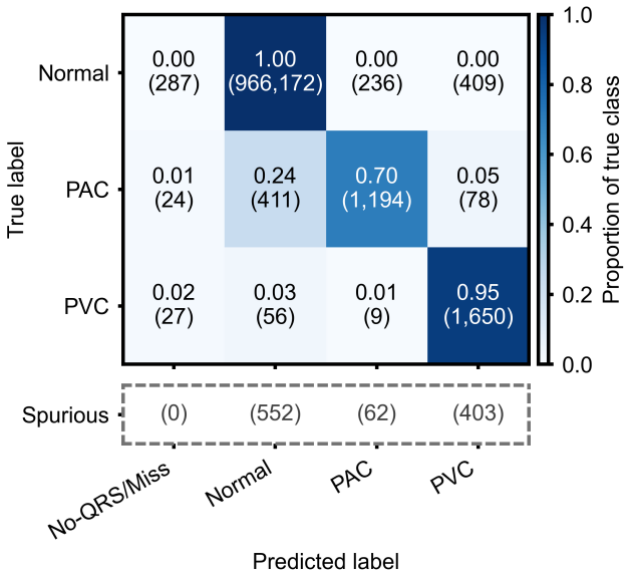

Figure 1. Beat-wise confusion matrix on the held-out test set (n=1430), row-normalised. The No-QRS/Miss column denotes annotated beats the model failed to detect; the Spurious row denotes false-positive QRS predictions.

Table 2. Model performance on the held-out test set measured on a beat level. The Overall row is the macro-average over the Normal, PAC, and PVC classes.

	F1	Precision	Recall
PAC	0.74	0.80	0.70
PVC	0.77	0.65	0.95
Normal	1.00	1.00	1.00
Overall	0.84	0.81	0.88

Bland–Altman analysis demonstrated good agreement between predicted and ground truth PAC counts, aggregated across the full ECG recording for each participant, with a mean bias of −0.14 and limits of agreement (LoA) -3.33 to 3.05 beats (Figure 2).

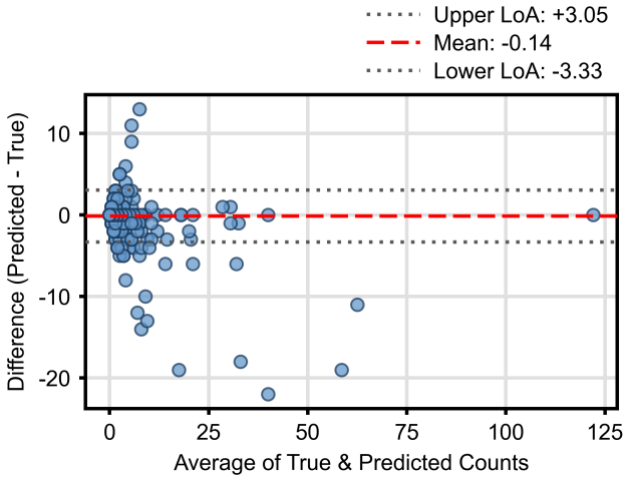

Figure 2. Bland-Altman analysis showing agreement between predicted and ground truth PAC counts per participant on the held-out test set.

4. Discussion

The CNN-Bi-LSTM U-Net achieved strong performance on the held-out test set, demonstrating feasibility of automated PAC detection in noisy single-lead exercise ECGs – to our knowledge, a setting not previously addressed in literature. As expected, PAC classification proved the most challenging, achieving the lowest F1 score across all classes. This is consistent with prior work [6,7,9,18] and reflects the inherent difficulty of distinguishing PACs from normal beats, particularly in single-lead recordings where morphological cues are limited [19] and exercise-induced noise creates artefacts imitating QRS complexes, leading to false positive detections [12] and thus impairing the crucial rhythm context.

Despite the lower PAC prevalence in the held-out test set relative to the training set, PAC precision remained high – a crucial property at population scale where false positives would systematically inflate PAC burden estimates. Bland–Altman analysis further demonstrated good participant-level agreement in PAC counts, indicating that aggregate burden estimates remain reliable for epidemiological use even when individual beat classifications are imperfect.

A major limitation is the small number of labelled ECGs. The training set ectopy-enrichment made annotation laborious, constraining dataset size and introducing potential label noise. Performance declined from cross-validation to evaluation despite strong dropout and regularisation. Distribution shift offers a partial explanation, but performance degradation is class-dependent: PVC degraded in precision while PAC degraded primarily in recall. A pure prevalence shift would typically degrade PAC precision rather than recall,

suggesting that PAC morphology and context – not only its distribution – differ between the training and test sets. Expanding the labelled pool or learning more generalisable representations via self-supervised learning is a priority for future work.

It is important to note that excluding the noisiest recordings and windows – together 1.1% of the test signal – and masking the Unknown class are likely to have inflated performance metrics slightly. We acknowledge this limitation but note that both are precautions intended for downstream epidemiological use, where unreliable predictions must be excluded.

This study demonstrates the feasibility of high-precision PAC detection in noisy single-lead exercise ECGs. The proposed approach enables scalable and reliable estimation of PAC burden, supporting its use for population-level phenotyping.

Acknowledgements

AD's research team is supported by a range of grants from the Wellcome Trust [223100/Z/21/Z, 227093/Z/23/Z], GSK, Boehringer Ingelheim, Google, National Institutes of Health's Oxford Cambridge Scholars Program, EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1), British Heart Foundation Centre of Research Excellence (RE/18/3/34214), Cancer Research UK, and funding administered by the Danish National Research Foundation in support of the Pioneer Centre for SMARTbiomed. Additional support was provided by a University of Oxford British Heart Foundation Centre of Research Excellence Fellowship awarded to SvD. AB is supported by the Oxford EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). RSW is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] D. Conen *et al.*, 'Premature atrial contractions in the general population', *Circulation*, vol. 126, no. 19, pp. 2302–2308, Nov. 2012.

[2] J. M. Farinha, D. Gupta, and G. Y. H. Lip, 'Frequent premature atrial contractions as a signalling marker of atrial cardiomyopathy, incident atrial fibrillation, and stroke', *Cardiovasc. Res.*, vol. 119, no. 2, pp. 429–439, Mar. 2023.

[3] J. C. L. Himmelreich *et al.*, 'Frequent premature atrial contractions are associated with atrial fibrillation, brain ischaemia, and mortality: a systematic review and meta-analysis', *Europace*, vol. 21, no. 5, pp. 698–707, May 2019.

[4] C. Carluccio *et al.*, 'Exercise stress testing in clinical cardiology: a practical guide to performance and interpretation', *J. Clin. Med.*, vol. 15, no. 4, p. 1656, Jan. 2026.

[5] M. Goswami, B. Boecking, and A. Dubrawski, 'Weak supervision for affordable modeling of electrocardiogram data', *AMIA Annu. Symp. Proc.*, vol. 2021, pp. 536–545, Feb. 2022.

[6] M. M. Farag, 'A tiny matched filter-based CNN for inter-patient ECG classification and arrhythmia detection at the edge', *Sensors*, vol. 23, no. 3, p. 1365, Jan. 2023.

[7] J. Hong, H. J. Li, C. C. Yang, C. L. Han, and J. C. Hsieh, 'A clinical study on atrial fibrillation, premature ventricular contraction, and premature atrial contraction screening based on an ECG deep learning model', *Appl. Soft Comput.*, vol. 126, p. 109213, Sep. 2022.

[8] M. Orini *et al.*, 'Premature atrial and ventricular contractions detected on wearable-format electrocardiograms and prediction of cardiovascular events', *Eur. Heart J. Digit. Health*, vol. 4, no. 2, p. 112, Mar. 2023.

[9] D. Kraft and P. Rumm, 'Automated detection of normal, atrial, and ventricular premature beats from single-lead ECG using convolutional neural networks', *Sensors (Basel)*, vol. 26, no. 2, p. 513, Jan. 2026.

[10] J. Zhang *et al.*, 'Opportunities and challenges of noise interference suppression algorithms for dynamic ECG signals in wearable devices: a review', *Measurement*, vol. 250, p. 117067, Jun. 2025.

[11] G. García-Isla, L. Mainardi, and V. D. A. Corino, 'A detector for premature atrial and ventricular complexes', *Front. Physiol.*, vol. 12, Jun. 2021.

[12] Z. F. Mohd Apandi, R. Ikeura, S. Hayakawa, and S. Tsutsumi, 'An analysis of the effects of noisy electrocardiogram signal on heartbeat detection performance', *Bioengineering (Basel)*, vol. 7, no. 2, p. 53, Jun. 2020.

[13] C. Sudlow *et al.*, 'UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age', *PLoS Med.*, vol. 12, no. 3, p. e1001779, Mar. 2015.

[14] S. V. Duijvenboden *et al.*, 'Prognostic significance of different ventricular ectopic burdens during submaximal exercise in asymptomatic UK Biobank subjects', *Circulation*, vol. 148, no. 24, pp. 1932–1944, Dec. 2023.

[15] 'ANSI/AAMI EC57:1998/(R)2003 - Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms'.

[16] W. Reklewski, K. Heryan, P. Augustyniak, and M. Miśkowicz, 'Multiplierless QRS detection algorithm for mobile ECG monitoring based on approximate computing', *Measurement*, vol. 249, p. 116969, May 2025.

[17] P. H. Charlton *et al.*, 'Detecting beats in the photoplethysmogram: benchmarking open-source algorithms', *Physiol. Meas.*, vol. 43, no. 8, p. 085007, Aug. 2022.

[18] J. Lim, D. Han, and K. H. Chon, 'QRS-centric beat-wise atrial fibrillation detection in ECG signals using deep neural networks', *Comput. Biol. Med.*, vol. 192, p. 110282, Jun. 2025.

[19] Y. Zhang and Y. Xia, 'Single-lead ECG arrhythmia classification based on peak-enhanced attention network and quality-aware GAN data augmentation framework', *Sensors*, vol. 26, no. 12, p. 3852, Jan. 2026.

Address for correspondence:

Anna M. Bator
Big Data Institute, Old Road Campus, Oxford, United Kingdom
anna.bator@exeter.ox.ac.uk